Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. **421,** 1155–1165 (2012) doi:10.1111/j.1365-2966.2011.20375.x

The detection of globular clusters in galaxies as a data mining problem

Massimo Brescia,[1] Stefano Cavuoti,[2] Maurizio Paolillo,[2⋆] Giuseppe Longo[2]† and Thomas Puzia[3]

[1]*INAF - Astronomical Observatory of Naples, via Moiariello 16, I-80131 Napoli, Italy*
[2]*Dipartimento di Scienze Fisiche, University Federico II, via Cinthia 6, I-80126 Napoli, Italy*
[3]*Department of Astronomy and Astrophysics, Pontificia Universidad Católica de Chile, 7820436 Macul, Santiago, Chile*

Accepted 2011 December 12. Received 2011 November 28; in original form 2011 October 10

ABSTRACT

We present an application of self-adaptive supervised learning classifiers derived from the machine learning paradigm to the identification of candidate globular clusters in deep, wide-field, single-band *Hubble Space Telescope* (*HST*) images. Several methods provided by the DAta Mining and Exploration (DAME) web application were tested and compared on the NGC 1399 *HST* data described by Paolillo and collaborators in a companion paper. The best results were obtained using a multilayer perceptron with quasi-Newton learning rule which achieved a classification accuracy of 98.3 per cent, with a completeness of 97.8 per cent and contamination of 1.6 per cent. An extensive set of experiments revealed that the use of accurate structural parameters (effective radius, central surface brightness) does improve the final result, but only by ∼5 per cent. It is also shown that the method is capable to retrieve also extreme sources (for instance, very extended objects) which are missed by more traditional approaches.

Key words: methods: data analysis – methods: statistical – globular clusters: general – galaxies: elliptical and lenticular, cD – galaxies: individual: NGC 1399.

1 INTRODUCTION

The need to effectively exploit the scientific information contained in current and future synoptic surveys has led to a renaissance of interest in the application of data mining (DM) methods to astronomical programmes. DM, in fact, seems to be among the few, if not the only, ways to cope with the complexity and size of existing and foreseen massive data sets such as those expected to be provided by the Large Synoptic Sky Telescope (LSST). The DM methods, however, are also very useful to capture the complexity of small data sets and, therefore, can be effectively used to tackle problems of much smaller scale. In this paper we used a variety of methods provided by the DAta Mining and Exploration Web Application REsource (DAMEWARE, http://dame.dsf.unina.it/beta_info.html) for the identification of globular clusters (GCs) in the galaxy NGC 1399 using single-band photometric data obtained with the *Hubble Space Telescope* (*HST*).

The identification and physical characterization of GC populations in external galaxies is of interest to many astrophysical fields: from cosmology to the evolution of star clusters and galaxies, to the formation and evolution of binary systems. The identification of GCs in external galaxies usually requires the use of wide-field, multiband photometry, since in galaxies located more than a few Mpc away they appear as unresolved sources in ground-based astronomical images and are thus hardly distinguishable from background galaxies which introduce significant contamination problems. For such reason, GCs are traditionally selected using methods based on their colours and luminosities. However, in order to minimize contamination and to measure GC properties such as sizes and structural parameters (core radius, concentration, etc.), high-resolution data are required as well which, for star clusters outside the Local Group, are available only through the use of space facilities (i.e. *HST*). Obtaining suitable *HST* data is however challenging in terms of observing time since the optimal data sets should be (i) deep, in order to sample the majority of the GC population and ensure the high signal-to-noise ratio (S/N) required to measure structural parameters (see e.g. Carlson & Holtzman 2001); (ii) with wide-field coverage, in order to minimize projection effects as well as to study the overall properties of the GC populations, which often differ from those inferred from observations of the central region of a galaxy only; and (iii) multiband, to effectively select GC based on colours.

It is apparent that, in order to reduce observing costs, it would be much more effective to use single-band *HST* data. Such approach however requires to carefully select the candidate GCs based on the available photometric and morphological parameters in order to avoid introducing biases in the final sample (see below). Here we intend to show that the use of properly tuned DM algorithms can yield very complete data sets with low contamination even

⋆E-mail: paolillo@na.infn.it
†Visiting Associate, California Institute of Technology, Pasadena, CA 91125, USA.

with single-band photometry, thus minimizing the observing time requirements and allowing us to extend such studies to larger areas and to the outskirts of nearby galaxies.

The paper is structured as follows. In Section 2 we describe the data used to test the various methods. In Section 3 we provide a short methodological and technical introduction to DAMEWARE and to some classification methods tested for the first time in an astronomical context. In Sections 4 and 5 we describe the results of the experiments and draw our conclusions.

2 THE DATA

The data set used in this experiment consists of wide-field *HST* observations of the giant elliptical NGC 1399 located at the heart of the Fornax cluster. This galaxy represents an ideal test case since, due to its distance ($D = 20.13 \pm 0.4$ Mpc, see Dunn & Jerjen 2006), it is possible to cover a large fraction of its GC system (out to $>5R_e$) with a limited number of observations. Furthermore, it is particularly challenging because, at this distance, GCs are only marginally resolved even by *HST*; in fact, at NGC 1399 distance 1 Advanced Camera for Surveys (ACS) pixel corresponds to 2.93 pc (1 arcsec = 97.7 pc). This data set was used by Paolillo et al. (2011) to study the connection between GCs and low-mass X-ray binaries (LMXB) and by Puzia et al. (in preparation) to study the structural properties of the GC population. We summarize below the main properties of the data set, and refer to these works for a more detailed description of the observations and of data analysis.

The optical data were taken with the *HST* ACS (programme GO-10129, PI: T. Puzia), in the *F*606*W* (broad *V* band) filter, with integration time of 2108 s for each field. The observations were arranged in a 3×3 ACS mosaic, and combined into a single image using the MULTIDRIZZLE routine (Koekemoer et al. 2002). The final scale of the images is 0.03 arcsec pixel^{-1}, providing Nyquist sampling of the ACS point spread function (PSF). The field of view (FOV) of the ACS mosaic covers 100 arcmin^2 (Fig. 1) extending out to a projected galactocentric distance of ∼55 kpc, i.e. $4.9r_e$ of

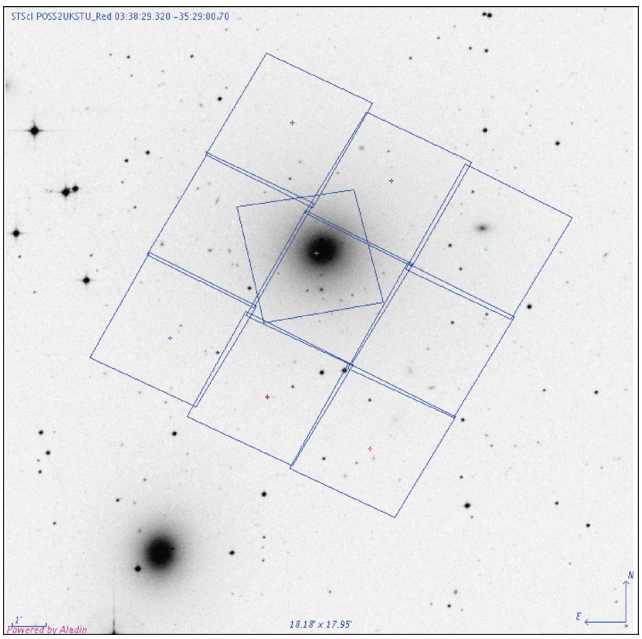

Figure 1. The FOV covered by the 3×3 *HST*/ACS mosaic in the *F*606*W* band. The central field, with a different orientation, shows the region covered by previous archival ACS observations in *g* and *z* bands.

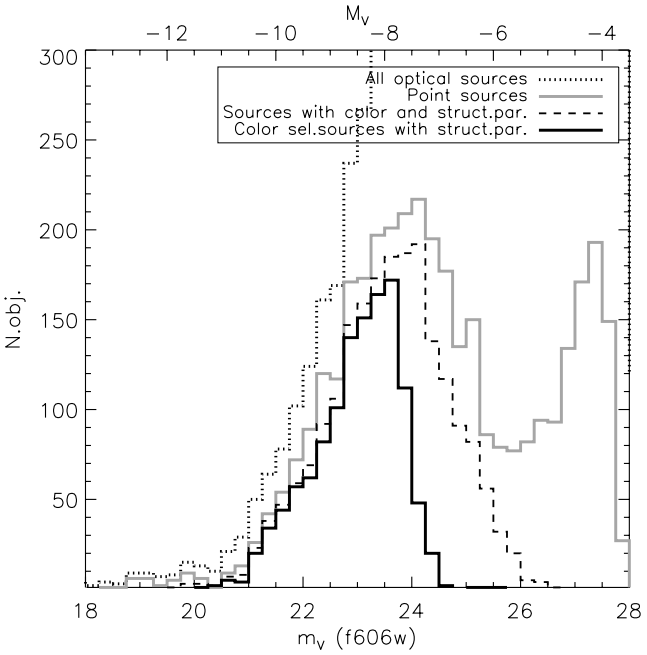

Figure 2. Luminosity distributions of all detected (dotted line) and point-like (e.g. with stellarity index >0.9, solid grey line) sourcesß within the *HST* FOV. Also shown are the two additional subsamples discussed in Sections 2 and 4: the KB composed of sources with both colour and structural parameters (dashed line), and the subset of bona fide colour-selected GCs based on Table 1 (solid black line).

the GC system ($\sim 5.7 r_e^{gal}$). The source catalogue was generated with SEXTRACTOR by imposing a minimum area of 20 pixels: it contains 12 915 sources and reaches 7σ detection at $m_V = 27.5$, i.e. 4 mag below the GC luminosity function (LF) turnover, thus allowing us to sample the entire GC population (see Fig. 2). The catalogue astrometric solution was registered to the USNO-B1.0 (United States Naval Observatory Catalog of Astrometric Standards) reference frame, obtaining a final accuracy of 0.2 arcsec rms.

For 4239 sources, we were able to measure structural parameters (which require very high S/N; see Carlson & Holtzman 2001; Puzia et al. 2012), fitting King surface brightness profile models with the GALFIT software (Peng et al. 2002) and deriving tidal, core, effective radii and central surface brightness values for each cluster. The accuracy of these measurements was estimated simulating several thousand artificial GCs with the MULTIKING code (available at: http://people.na.infn.it/paolillo/Software.html) specifically written to account for ACS field distortion, PSF variation, dithering pattern (Paolillo et al. 2011; Puzia et al., in preparation).

The NGC 1399 region covered by our mosaic lacks colour information for all *HST* *F*606*W* sources. In this paper we shall therefore make use of two ancillary multiwavelength data sets: archival *HST* $g - z$ observations (Kundu et al. 2005), which cover the very central region of the galaxy (10 per cent of the sample, see Fig. 1), and $C - T1$ ground-based photometry from Bassino et al. (2006), covering the whole mosaic. The latter is only available for ∼14 per cent of our sources, and due to background light contamination it is very incomplete in the proximity of the galaxy centre. In total, 2740 sources of the catalogue have multiband (either $g - z$ or $C - T1$) photometry.

Finally, the subsample of sources used to build our knowledge base (KB, see Section 3) to train the DM algorithms is composed of the 2100 sources with all photometric and morphological

Table 1. Photometric selection criteria for GC candidates.

	Colour cut	Magnitude cut
Ground-based data	$1.0 \leq C - T1 < 2.2$	$T1 < 23$
HST data	$1.3 \leq g - z < 2.5$	$z < 22.5$

information: isophotal magnitude, Kron radius, aperture magnitudes within a 2, 6 and 20 pixels (corresponding to 0.06, 0.18 and 0.6 arcsec) diameter, ellipticity, position angle, full width at half-maximum (FWHM), SExtractor stellarity index, King's tidal and core radii, effective radii, central surface brightness and either $g - z$ or $C - T1$ colour. The magnitude distribution of such subsample is shown in Fig. 2 as a dashed line.

The typical choice to select GCs based on multiband photometry would be to adopt the magnitude and colour cuts reported in Table 1, and highlighted in Fig. 3 with a dashed line; the magnitude limit $z < 22.5$ does not exploit the full depth of the *HST* data but is adopted to be consistent with the $T1 < 23$ limit used for the ground-based colours, thus ensuring a uniform limit across the whole FOV. In the following, we thus assume that bona fide GCs are represented by such sources, in order to explore how well different selection methods based on single-band photometry are able to retrieve the correct population of objects. The $F606W$ magnitude distribution of colour-selected GCs is shown in Fig. 2 as a black solid line.

3 SOME CONSIDERATIONS ON DATA MINING

DAMEWARE (Brescia et al. 2010) is one of the main products made available through the DAta Mining and Exploration (DAME) programme collaboration. It provides a web browser based front-end, able to configure DM experiments on massive data sets and to execute them on a distributed computing infrastructure (cloud/grid hybrid platform). DAMEWARE offers the possibility to access different DM functionalities (supervised classification, regression and clustering) implemented with different methods [traditional multilayer perceptrons (MLP), support vector machines (SVM), etc.]. Even though specifically designed to deal with massive data sets, DAMEWARE can also be used on small ones. It needs however to be taken into account that, due to the poor coverage of the parameter space by the KB, DM on small data sets requires special care. In what follows we shall outline the main strategy behind our procedure.

The problem tackled in this work is a typical supervised classification task and therefore, while referring the reader to Duda, Hart & Storck (2004) and Bishop (1995) for a general introduction to DM, we shall shortly summarize some aspects which are relevant to the experiments described in the next paragraph.

First of all, it needs to be kept in mind that in the DM practice, there is no way to a priori select the algorithm which offers the best performances for a given task and that therefore a number of trial and error experiments must be performed in order to identify the method with the best performances. From a logical point of view, effective supervised classification is based on the following steps.

(i) To select and create the data parameter space, i.e. to create the data input patterns (or features) to be submitted to the classifiers. It is important in this phase to build homogeneous patterns, i.e. with each pattern having the same type and number of parameters.

(ii) To prepare the data sets which are needed for the different experiment steps: training, validation and test sets (the data set must include also target values for each input pattern, i.e. the desired output values, coming from any available knowledge source), by splitting the KB into variable subsets to be submitted at each phase.

(iii) To analyse and select classification model, based on theoretical principles and on the user experience about the content of the KB.

(iv) To perform complete sequences of experiments with all model candidates and compare their results in terms of training error, learning robustness, output correctness (as defined below); this phase might also require a pruning of the parameter space.

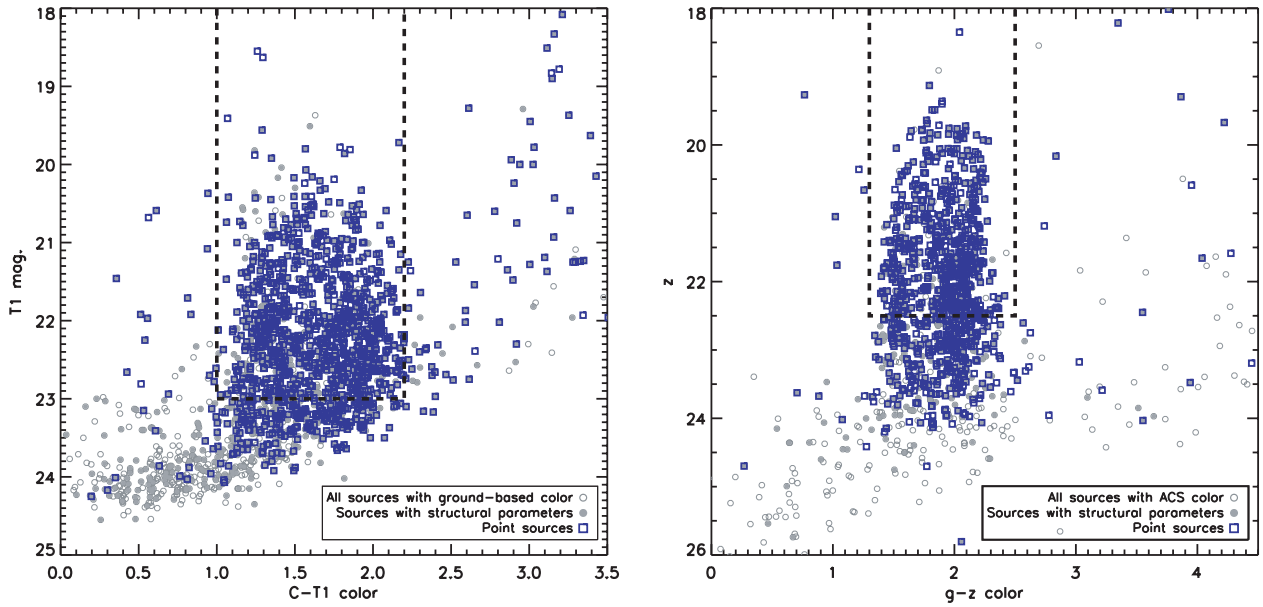

Figure 3. Colour–magnitude diagrams using $C - T1$ ground-based (left-hand panel) and $g - z$ *HST* photometry (right-hand panel). Ground-based photometry covers the whole FOV of our ACS mosaic, while *HST* colours are limited to the central ACS field ($\sim$200 $\times$ 200 arcsec^2, Fig. 1). Open grey dots represent all sources in colour catalogues while solid ones refer to the subsample with both colour and structural parameters that represents our KB. Blue squares mark point-like sources, i.e. sources with stellarity index >0.9, while the dashed line highlights the parameter space (Table 1) used to select bona fide GC.

(v) Finally, to identify the best model which will then be adopted as the final classifier to be applied to the entire data set.

Optionally (either because some methods do not require it or simply as a user choice) a validation procedure may be introduced. Validation is the process of checking whether the classifier meets some criterion of generality when dealing with unseen data in order to avoid overfitting or to stop the training on the base of an 'objective' criterion. Here 'objective' implies a criterion which is not based on the same data used for the training procedure. Obviously, validation requires an additional data set which can be prepared by the user directly or in an automatic fashion.

When the training set is of limited size – such as the one used in this paper – it is almost unavoidable to adopt a 'subset validation' procedure. This implies the partitioning of a sample of data into subsets, such that the analysis is initially performed on a single subset, while the other subset(s) are retained for subsequent use in confirming and validating the initial analysis. In practice, the data sample is divided into N subsets, some of which are used for the training phase (training set), while the others are employed, as validation sets, to compare the model prediction capability. By varying the value of N (different splitting of the data sets) it is possible to evaluate the prediction accuracy of the trained model (Kotsiantis 2007).

The so called k-fold cross-validation divides the whole data set into K subsets, each of them is alternately excluded from the validation set. In practice, all data are used for the training and test phases in an independent way. In this case, we obtain K classifiers ($2 \leq K \leq n$) whose outputs can be used to obtain a mean evaluation. The downside of this method is that it is very expensive in terms of computing time in the case of massive data sets.

As it was briefly mentioned, in a supervised machine learning scheme, the training is done by means of a mechanism in which the model output is compared with the desired target output for each input pattern, allowing us to define a training error. The choice of the metric function used for the comparison (which defines the training error) determines the evaluation criteria and the learning rule of the model. Different error evaluation metrics exist in literature, depending on the problem complexity to be solved. In our experiments we used several methods. The most common metric is the mean square error (MSE) of the difference between model and target outputs. Supervised neural networks that use MSE cost function can use formal statistical methods to determine the confidence of the trained model (Yang & Shanna 1991), while the MSE computed on a validation set can be used as an estimate of the variance. This value can then be used to calculate the confidence interval of the output of the network assuming a normal distribution. A confidence analysis made in this way is statistically significant as long as the output probability distribution remains the same and the network is not modified.

By assigning a softmax activation function (Bishop 1995) on the output layer of the neural network (or a softmax component in a component-based neural network) for categorical target variables, the outputs can be interpreted as posterior probabilities (Sutton & Barto 1998). This is very useful in classification as it gives a certainty measure on classifications.

Many supervised models also support the use of the cross-entropy error function for addressing classification problems in a consistent statistical fashion (Rubinstein & Kroese 2004).

The cross-entropy method consists of two phases:

(1) generate a random data sample (trajectories, vectors, etc.) according to a specified mechanism;

(2) update the parameters of the random mechanism based on the data to produce a 'better' sample in the next iteration.

In practice, a data model is created based on the training set, and its cross-entropy is measured on a test set to assess how accurate the model is in predicting the test data. In practice, the method compares two probability distributions, p the true distribution of data in any corpus and q the distribution of data as predicted by the model. Since the true distribution is unknown, cross-entropy cannot be directly calculated; an estimate of cross-entropy is calculated using the following formula:

$$H(T, q) = -\sum_{i=1}^{N} \frac{1}{N} \log_2 q(x_i),$$

where T is the chosen training set, corresponding to the above-mentioned true distribution p, N is the number of objects in the test set, and $q(x)$ is the probability of the event x estimated from the training set.

Due to the supervised nature of the classification task, the system performance can be measured by means of a test set during the testing procedure, in which unseen data are given to the system to be labelled. The overall performance thus integrates information about the classification accuracy (i.e. in terms of output correctness). Moreover, the results obtained from the unseen data are also important to evaluate the learning robustness, i.e. the generalization capability of the network in presence of data samples never used during the training phase. However, when a data set is unbalanced (i.e. when the number of samples in different classes varies greatly) the error rate of a classifier is not representative of the true performance of the classifier itself.

For the specific problem addressed in this paper, we used five among the different classification methods available in DAMEWARE. Namely multilayer perceptron trained by back propagation (MLP-BP), SVM, genetic algorithm model experiment (GAME), MLP with genetic algorithms (MLPGA) and multilayer perceptron trained by quasi-Newton (MLPQNA). MLP-BP and SVM have already been described several times in the astronomical literature and therefore we refer the reader to Bishop (1995) and Chang & Lin (2011). For what the other methods are concerned, since they are used for the first time in an astronomical context, we shall provide some further details.

3.1 The multilayer perceptron trained by genetic algorithms

Genetic algorithms (GAs) are computational methods inspired by Darwin's evolutionary mechanism (Holland 1975). GAs are particularly powerful in solving problems where the solution space is not well defined. When they are embedded into an MLP network, the resulting learning algorithm (named MLPGA model) consists mainly in the cyclic exploration of the parameter space aimed at discovering the best solution (Meng & Fan 2009).

In a GA, each element of a population (i.e. each data point) is called chromosome and is composed by a set of genes (features) that represents its deoxyribonucleic acid (DNA). From a more traditional point of view, each DNA can be therefore considered as a possible solution to the problem. The starting point of the method consists in the random generation of a population of chromosomes, for example by using normal or uniform statistical distributions. Then the method proceeds by cyclic variation and combination of the initial population, modifying their DNAs (neuron weights) according to the standard feed-forward MLP calculations on input

patterns. The final goal is to find the best population (best problem solution), where 'best' is defined according to some fitness criterion.

In other words, at each evolutionary step (backward phase of the MLPGA model), the output chromosomes are obtained by applying several genetic operators to the input population and by evaluating through a specific fitness function the goodness of the newly generated population. The fitness function provides a method to discard the worst chromosomes from the population, thus allowing only the best candidates to evolve to the next generation (similarly to what happens in natural selection). The entire cycle is iterated until the chromosome with the desired fitness is found (i.e. the best solution to the classification problem). The training error calculation follows the MSE criterion.

3.2 The genetic algorithm model experiment

As it was briefly mentioned above, this machine learning model arises from an original customization, made by DAME group, of the standard generalized GA model. All basic theoretical aspects for a generic GA have already been presented in the MLPGA section. The idea behind the GAME model is to create a special fitness function, based on a polynomial expansion approximation, able to perform supervised adaptive learning on massive data sets. The analytical expression used to solve classification problem is the trigonometric series expansion of each input pattern features, compared with the corresponding known pattern target value. Then the whole error (MSE), which is the fitness function, is calculated at each cycle for all input patterns, and the population of genetic chromosomes is updated according the classical genetic operators (crossover and mutation). This loop ends when the minimum error is found (below a chosen error threshold) or if the maximum number of iteration is reached.

3.3 The multilayer perceptron trained by quasi-Newton rule

Quasi-Newton algorithms (QNAs) are variable metric methods for finding local maxima and minima of functions (Davidon 1991). The model based on this learning rule and on the MLP network topology is then called MLPQNA. QNAs are based on Newton's method to find the stationary (i.e. the zero gradient) point of a function. Newton's method assumes that the function can be considered as quadratic in a narrow region around the optimum and uses the first and second derivatives (gradient and Hessian) to find the stationary point. In QNA, the Hessian matrix of second derivatives of the function to be minimized does not need to be computed and can be derived by analysing successive gradient vectors. QNA is a generalization of the secant method to find the root of the first derivative for multidimensional problems. In multidimensions, the secant equation is underdetermined, and quasi-Newton methods differ in how they constrain the solution, typically by adding a simple low-rank update to the current estimate of the Hessian. Since, as it will be shown, this model performed the best in the GC classification problem discussed in this paper, we shall discuss it in more detail.

In DAMEWARE, the quasi-Newton method has been implemented by following the known L-BFGS (Limited memory–Broyden–Fletcher–Goldfarb–Shanno) algorithm (Byrd, Nocedal & Schnabel 1994). The QNA is an optimization of Newton-based learning rule, also because, as described below, the implementation is based on a statistical approximation of the Hessian by a cyclic gradient calculation, that is at the base of back propagation method. By using a local square approximation of the error function, we can obtain an expression for the minimum position. The gradient in every point w is in fact given by

$$g = \nabla E = \mathbf{H} \times (\boldsymbol{w} - \boldsymbol{w}^*), \tag{1}$$

where $\boldsymbol{w}^*$ corresponds to the minimum of the error function, which satisfies the condition:

$$\boldsymbol{w}^* = \boldsymbol{w} - \mathbf{H}^{-1} \times g. \tag{2}$$

The vector $-\mathbf{H}^{-1}g$ is known as Newton direction and it is the base for a variety of optimization strategies, such as the QNA which instead of calculating the $\mathbf{H}$ matrix and then its inverse uses a series of intermediate steps of lower computational cost to generate a sequence of matrices which are more accurate approximations of $\mathbf{H}^{-1}$. These matrices are computed using only information related to the first derivative of the error function.

The Newton direction can be used in a line search (optimization problem) method when the Hessian matrix $\mathbf{H}$ is positive definite, because under such requirement it is a descent direction. When the Hessian is not positive definite, the Newton direction may not be defined, because its inverse matrix may not exist. However, in addition, also when it is definite, it may not satisfy the descent trend. In particular, the main drawback of the Newton direction is the need for the exact Hessian matrix formulation, which is described in more detail in Appendix A.

As a matter of fact, this method was designed to optimize the functions of a number of arguments (hundreds to thousands), because in this case it is worth having an increased iteration number due to the lower approximation precision because the overheads become much lower. This is particularly useful in astrophysical DM problems, where usually the parameter space is dimensionally huge and is often afflicted with a low S/N.

4 RESULTS

As discussed in Section 1, the purpose of this work was to implement an alternative, DM-based, method to select GCs in single-band *HST* images, thus saving the observing time needed to obtain complete sets of multiband data. In this section, we shortly summarize the results of the series of (numerical) 'experiments' which were performed to determine the best model and the best combination of features, while in the next section we discuss the overall properties of the sample obtained with the DM algorithms, in comparison with traditional selection methods.

Terms like completeness, contamination, accuracy, etc. are differently defined by astronomers and 'data miners'. In what follows we use the following definitions. Classification accuracy: fraction of patterns (objects) which are correctly classified (either GCs or non-GCs) with respect to the total number of objects in the sample; completeness: fraction of GCs which are correctly classified as such; contamination: fraction of non-GC objects which are erroneously classified as GCs.

All experiments were performed on the KB sample presented in Section 2, assuming that bona fide GCs are represented by sources selected according to the colour cuts in Table 1. We used as features the following quantities:

(i) the isophotal magnitude (feature 1);

(ii) three aperture magnitudes (features 2–4) obtained through circular apertures of radii 2, 6 and 20 arcsec, respectively;

(iii) the Kron radius, the ellipticity and the FWHM of the image (features 5–7);

Table 2. Summary of the performances (in percentage) of the five classifiers. For each entry, the first line refers to the classification accuracy, while the second and third refer to completeness and contamination, respectively. Values in bold highlight the experiments yielding the best results.

Type of experiment	Missing features	Figure of merit	MLPQNA	GAME	SVM	MLPBP	MLPGA
Complete patterns	–						
		class.accuracy	**98.3**	82.1	90.5	59.9	66.2
		completeness	**97.8**	73.3	89.1	54.1	61.4
		contamination	**1.8**	18.7	7.7	42.2	35.1
No par. 11	11						
		class.accuracy	**98.0**	81.9	90.5	59.0	62.4
		completeness	**97.6**	79.3	88.9	56.1	62.2
		contamination	**1.6**	19.6	7.9	43.1	38.8
Only optical	8, 9, 10, 11						
		class.accuracy	93.9	86.4	90.9	70.3	76.2
		completeness	91.4	78.9	88.7	54.0	65.1
		contamination	5.9	13.9	8.0	33.2	24.6
Mixed	5, 8, 9, 10, 11						
		class.accuracy	94.7	86.7	89.1	68.6	71.5
		completeness	92.3	81.5	88.6	52.8	63.8
		contamination	5.0	16.6	8.1	37.6	30.1

(iv) the structural parameters (features 8–11) which are, respectively, the central surface brightness, the core radius, the effective radius and the tidal radius.

By making an exhaustive pruning test on all 11 data set parameters, with the five machine learning models previously introduced, we collected a total of 425 experiments (85 per model). The details of the experiment set-up can be found in Appendix B.

Table 2 summarizes the most relevant results: in terms of classification accuracy and completeness, the best results (98.3 and 97.8 per cent, respectively) are obtained by MLPQNA using all parameters; using all available features but the number 11 (the tidal radius), we obtain marginally worse results, as can be expected given the high noise present in this last parameter, which is affected by the large background due to the host galaxy light. In terms of contamination, comparable results ($\lesssim$2 per cent) are obtained with the same model both with or without feature 11. We point out that since the experiment without feature 11 provides results comparable to the one using all features, but requires less information and is less computationally demanding, we consider the latter to be the case providing the highest overall performance, as usually done in DM experiments. In other words, the experiment without feature 11 represents the best compromise between required overall performance and complexity of the KB.

The best result obtained without using the structural parameters is 93.9 per cent (classification accuracy), thus indicating that the availability of detailed structural parameters does indeed help to improve the results, but only by $\sim$5 per cent. Moreover, the pruning in the mixed cases (by excluding some structural and optical features) revealed a similar behaviour in all models, in terms of quantity of correlated information introduced by individual features in the patterns. Five optical features (namely the isophotal and aperture magnitudes and the FWHM of the image) were recognized as the most relevant by all models. Among the structural parameters, the central surface brightness and the core radius were recognized as relevant by all models but the SVM and MLPGA models. In all other cases, other residual optical and structural parameters were evaluated low carriers of correlated information.

It is worth pointing out that the performances, in terms of completeness and contamination, quoted above are all derived with respect to the test sample (and thus ultimately from the KB), and do not include possible biases affecting the KB itself. Such biases will be propagated to the final sample by any DM algorithm, as these rely on the assumption that the KB is a fair and complete representation the 'real' population that we want to identify. Thus, if the KB is severely incomplete or contaminated, this is a separate issue that has to be addressed in the training sample selection phase.

5 DISCUSSION

In order to test the effectiveness of our method, we need to compare its performances with those offered by more traditional approaches. For homogeneity (same data set) we shall use as template the method discussed in Paolillo et al. (2011) which used a selection criterion based on magnitude and morphology. Fig. 2 shows that sources with SExtractor stellarity index >0.9 (grey solid line) are distributed as the GC LF down to $m_V = 26$, while at fainter magnitudes background unresolved sources dominate the overall sample. Based on these considerations, Paolillo et al. choose as GC candidates sources having stellarity index >0.9 and $m_V < 26$ mag. Clearly a more sophisticated selection process, based on complex combinations of photometric and structural parameters (see for instance Puzia et al., in preparation), could be adopted, but any such approach requires anyway extensive testing to verify what biases are introduced in the final sample and it is not clear how such biases can be evaluated and corrected for without the availability of additional data (e.g. more uniform colour coverage or random background fields to compare with).

From Fig. 3, it can be seen that although the use of the stellarity and magnitude criteria effectively selects the bulk of the colour-selected GC population, there are sources consistent with GC colours, which are missed by this approach; on the other hand, this subsample includes many objects outside the allowed colour range. We can calculate the level of completeness and contamination resulting from the simple approach of Paolillo et al. (2011), as done in Section 4 for the DM methods. We derive two different estimates (i) for the central region covered by the more accurate *g* and *z HST* photometry and (ii) for the entire field covered by the ground-based *C* and *T*1 data. Within the central region, 92 per cent

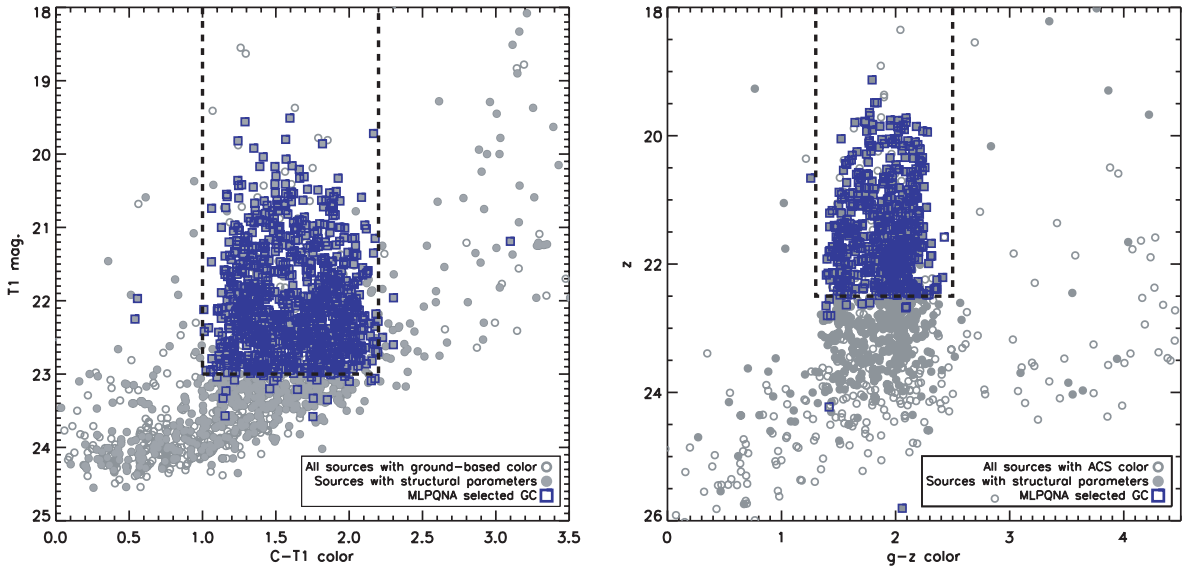

Figure 4. Same as Fig. 3 showing the colour distribution of the MLPQNA-selected sample. The MLPQNA sample (blue squares) reproduces the properties of the colour-selected GC population (i.e. the KB) with much less contaminants than e.g. the point-like population shown in Fig. 3.

of our GC candidates (within $m_V < 26$ by definition) are consistent with the $1.3 \leq g - z < 2.5$ colour cut and $z < 22.5$. Using the $C - T1$ photometry instead, which extends over the whole *HST* mosaic, we find that 82 per cent of the GC candidates are consistent with the $1.0 \leq C - T1 < 2.2$ colour and $T1 < 23$ mag cuts. On the other hand, ∼4 and ∼9 per cent of the GC candidates have, respectively, $g - z$ and $C - T1$ colours outside the allowed range as given in Table 1.

When these numbers are compared with those presented in Table 1, we see that the MLPQNA outperforms the simpler approach used by Paolillo et al. (2011) both in the central region and across the whole field, in the sense that it results in higher completeness, retrieving a larger fraction of the colour-selected sources using only single-band photometry. GAME and SVM may still perform better in the galaxy outskirts, although in the galaxy centre they are slightly less accurate. In terms of contamination, the MLPQNA again performs better than the Paolillo et al. (2011) approach, yielding <2 per cent spurious sources in the two best experiments (complete patterns and no par.11). The other MLPQNA experiments and all SVM cases are still competitive in the galaxy outskirts.

The performance of the MLPQNA method is better understood looking at the colour–magnitude plot shown in Fig. 4. The MLPQNA sample reproduces the properties of the colour-selected GC population with much less contaminants, than e.g. the point-like population shown in Fig. 3, and less outliers. In Fig. 5, we show the luminosity distribution of the MLPQNA sample: the MLPQNA approach (dashed red line) is able to retrieve almost the entirety of the colour-selected GC population (solid black line). We point out that the luminosity limit at $m_V \sim 24$ is due to the magnitude threshold imposed on the colour-selected sample (Table 1) in order to get a uniform limit across the whole colour range (Fig. 4) and FOV, and is thus not an intrinsic feature of the GC LF which extends down to $m_V \gtrsim 26$ mag.

A detailed investigation of the properties of the spurious sources is difficult since the strength of DM algorithms is to detect *hidden* correlations among the parameters, and use them to classify unknown sources; this however means that such correlations are hard to identify through a simple (and low-dimensional) view of the source distribution in the parameter space. In our specific case we

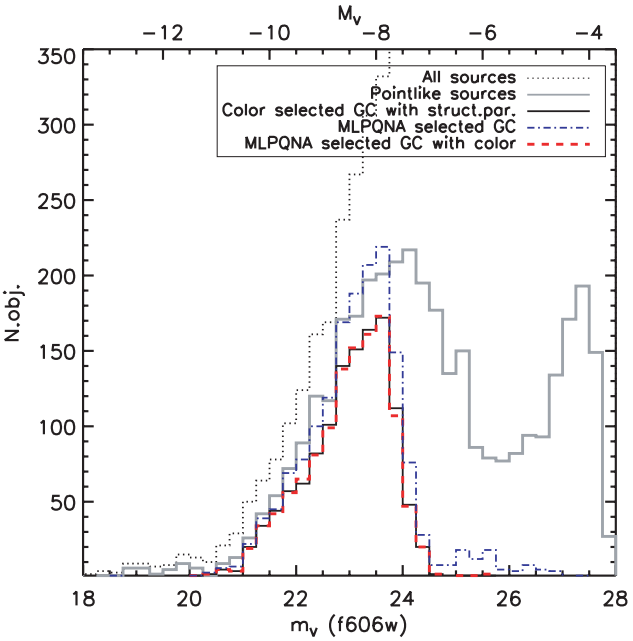

Figure 5. Same as Fig. 2 but for the MLPQNA-selected samples. The MLPQNA approach (dashed red line) is able to retrieve almost the entirety of the colour-selected GC population (solid black line); applying the same algorithm to all sources with structural parameters (but no colour, blue dot–dashed line), we can thus retrieve many more objects than available in the colour-selected subsample, sharing the same luminosity distribution of the latter population.

found that most contaminants are indeed GCs which fail the colour–magnitude classification technique by only one criterion (e.g. they lie just outside the chosen colour or magnitude range, see Fig. 4). It is thus unsurprising that the MLPQNA identifies such sources as GCs as all other parameters obey the correlations identified in the training phase. A few more extreme objects are found to be affected by photometric or structural problems in the data, such as the overlap with a nearby source which may introduce severe contamination in low-resolution ground-based data, or a position close to the chip

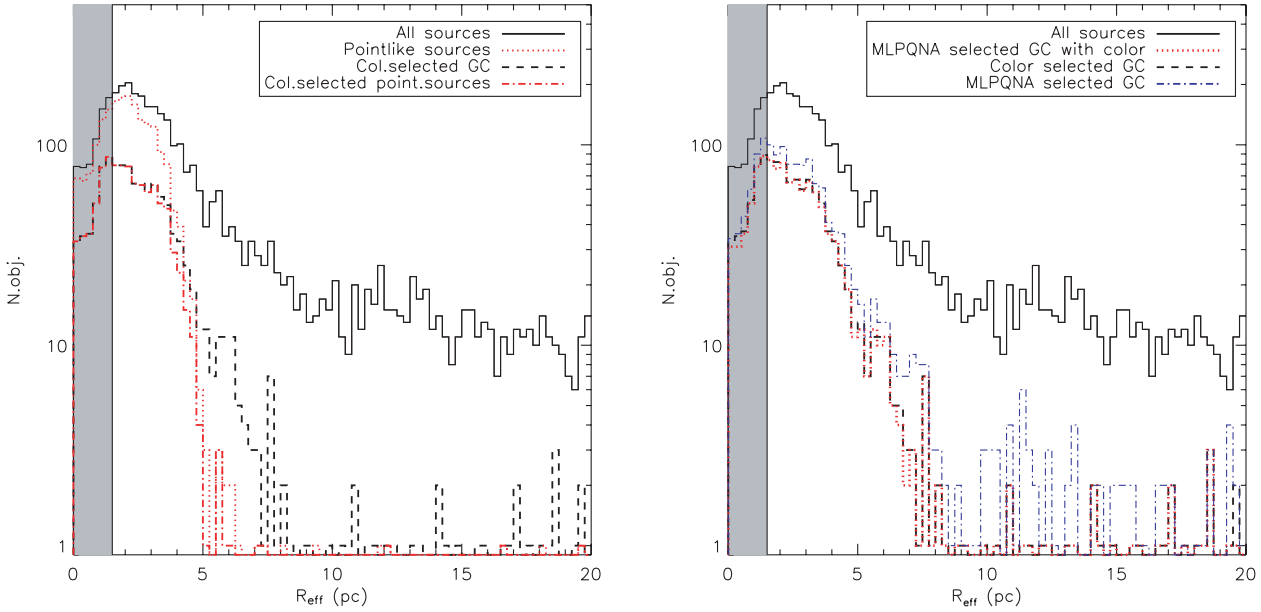

Figure 6. Left-hand panel: half-light radius distribution for the entire ACS optical catalogue (solid line), compared to Paolillo et al. (2011) GC candidates, i.e. point-like sources with $m_V < 26$ (dotted line). Restricting the sample to colour-confirmed GCs (dashed and dot–dashed lines) shows that the Paolillo et al. (2011) selection criteria miss very extended GCs with $R_{\rm eff} > 5$ pc. The shaded region highlights the region where our size measurement is poorly constrained (see Paolillo et al. 2011; Puzia et al., in preparation). Right-hand panel: same as left-hand panel but for the MLPQNA-selected samples. The MLPQNA-selected sample (dotted red line) reproduces the size distribution of the colour-selected GC population (dashed black line), thus avoiding the size biases resulting from the simpler Paolillo et al. (2011) selection criteria; the same is true when applying the MLPQNA algorithm to the larger subsample with structural parameters (blue dot–dashed line).

gap in *HST* data. A few can also be expected to be foreground stars which are misclassified due to the small angular size of some GCs in the training sample, lying at the resolution limit of *HST* data (grey region in Fig. 6).

An additional advantage in the use of DM techniques can be seen comparing the structural parameters of point-like sources with $m_V < 26$ mag with those of the colour-selected subsample (Fig. 6, left-hand panel): we find that the Paolillo et al. (2011) selection criteria miss extended sources with $R_{\rm eff} \gtrsim 5$ pc, as it can be expected given the compactness requirement (stellarity index >0.9). The right-hand panel of the same figure shows that the MLPQNA methods are instead able to retrieve also the most extended GCs. While some of these extended sources may be background galaxies, we point out that the most extended GCs, such as the Galactic GC ω-Cen, do fall in this range. In fact, we used the subset of GC confirmed through radial velocity (RV) measurements (Dirsch et al. 2004) to verify that a significant fraction ($\sim$10 per cent) of the NGC 1399 GC population has $R_{\rm eff} \gtrsim 5$ pc, and that the size distribution of this subsample is statistically indistinguishable from both the colour- and MLPQNA-selected populations. Obviously we cannot confirm that *all* extended sources are genuine GCs but, as already discussed in Section 4, we emphasize that the performance of the method has to be evaluated only by its ability to retrieve the same sources included in the training sample, i.e. the colour-selected GCs in our experiment.

Applying the same algorithm to the larger ensemble of sources with structural parameters (but no colour information), we are now able to retrieve more objects than available in the colour-selected subsample, sharing very similar properties to the latter population. The population of MLPQNA-selected GCs identified within the whole population is shown in Figs 5 and 6 (right-hand panel) as a dot–dashed line. In our specific test case (e.g. NGC 1399), this method allows us to identify $\sim$30 per cent more GCs than relying the subsample of sources with colour; this larger sample closely follows the GC LF down to the magnitude limit imposed by the colour selection, as well as the structural properties of the bona fide GC population. Thus, the gain with respect to other selection techniques is in the ability to retrieve a larger population with well-defined properties, at lower observational cost. In other programmes the gain can be much larger: for instance, in cases of large surveys where DM algorithms can be trained on a KB consisting of a limited number of multiband observations covering only a small fraction of the FOV, the trained algorithm will then allow us to extract statistically equivalent samples from the entire survey.

Finally, we note that each experiment was not really time consuming. It was executed on a common desktop multicore PC in a multithreading environment, resulting in about 3600 s (1 h) of duration for the training phase in the worst case (i.e. on the whole data set patterns with all 11 features). The test phase is instantaneous, since the trained network acts like a one-shot function. Of course, the complexity and indeed the execution time depends in a quadratic form on the data set dimension. But in the case of small data sets, like the present one, this is not an issue. Besides computing time, the relevant result is that the proposed MLPQNA model revealed a strong performance also in the case of small data sets where, as known, machine learning method performances are usually degrading, due to the limited size of the training samples. This is demonstrated by the poorer results obtained by other methods, shown in Table 2, which usually perform significantly better on larger data sets.

6 CONCLUSION

We performed an experiment showing that the use of DM techniques on small data sets allows us to solve complex astronomical problems such as the selection of GC candidates in external galaxies, from single-band images, provided that a subsample of sources can be used to train the DM algorithm. Since such methods do not assume any a priori model of the population, we are looking for, they allow us to retrieve samples which share the same properties of the training sample and are affected by less biases than results using simpler selection techniques.

In principle, we could use more refined approaches than those tested here, such as the use of RV measurements to improve the reliability of the KB, but any such approach would require the availability of additional data, i.e., in this particular case, spectroscopic observations. Such type of data are difficult to obtain and expensive in terms of observing time, thus justifying the DM methods proposed in this work. Obviously, in some instances these data could already be available in the archives, as for the NGC 1399 case where they have been used to verify some of our results (Section 5).

As a closing remark, we can safely state that, in the emerging scenario of the data-driven science, a DM-based approach to data analysis and interpretation seems to provide a large competitive edge over classical methods in particular for what concerns the ability to recognize patterns and derive correlations in high dimensionality data set that are not easily handled by human perception.

ACKNOWLEDGMENTS

The authors wish to thank the whole DAMEWARE working group, whose huge efforts made the DM facility available to the scientific community. MP acknowledges support from PRIN-INAF 2009, and thanks the ASI Science Data Center (ASDC) for support and hospitality. GL wishes to thank Professor G. S. Djorgovski and the whole Department of Astronomy at the California Institute of Technology in Pasadena, for hospitality. We also thank the anonymous referee for useful suggestions and comments.

REFERENCES

Bassino L. P., Faifer F. R., Forte J. C., Dirsch B., Richtler T., Geisler D., Schuberth Y., 2006, A&A, 451, 789
Bishop C. M., 1995, Neural Networks for Pattern Recognition. Oxford Univ. Press, Oxford
Brescia M. et al., 2010, preprint (arXiv:1010.4843B)
Broyden C. G., 1970, J. Inst. Math. Appl., 6, 76
Byrd R. H., Nocedal J., Schnabel R. B., 1994, Math. Program., 63, 129
Carlson M. N., Holtzman J. A., 2001, PASP, 113, 1522
Chang C.-C., Lin C.-J., 2011, ACM Trans. Intelligent Syst. Technol., 2, 27
Davidon W. C., 1991, SIAM J. Optim., 1, 1
Dirsch B. et al., 2004, AJ, 127, 2114
Duda R. O., Hart P. D., Storck D. G., 2004, Pattern Classification, 2nd edn. Wiley, New York
Dunn L. P., Jerjen H., 2006, AJ, 132, 1384
Fletcher R., 1970, Comput. J., 13, 317
Goldfarb D., 1970, Math. Comput., 24, 23
Holland J. H., 1975, Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor
Koekemoer A. M., Fruchter A. S., Hook R., Hack W., 2002, HST Calibration Workshop, 337
Kotsiantis S. B., 2007, Informatica, 31, 249
Kundu A. et al., 2005, ApJ, 634, L41
Meng J. E., Fan L., 2009, Proc. 21st Annu. Int. Conf., Chinese Control and Decision Conference. IEEE Press, Piscataway, NJ
Paolillo M., Puzia T. H., Goudfrooij P., Zepf S. E., Maccarone T. J., Kundu A., Fabbiano G., Angelini L., 2011, ApJ, 736, 90
Peng C. Y., Ho L. C., Impey C. D., Rix H.-W., 2002, AJ, 124, 266
Rubinstein R. Y., Kroese D. P., 2004, The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte Carlo Simulation, and Machine Learning. Springer-Verlag, New York
Shanno D. F., 1970, Math. Comput., 24, 647
Sutton R. S., Barto A. G., 1998, Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA
Yang X., Shamma S. A., 1991, Biological Cybernetics, 65, 171
Zhu C., Byrd R. H., Lu P., Nocedal J., 1997, ACM Trans. Math. Softw., 23, 550

APPENDIX A: QUASI-NEWTON LEARNING RULE

Quasi-Newton direction search methods provide a very useful alternative in that they do not require a precise calculation of the Hessian. In place of the Hessian matrix $\mathbf{H}_k$, they use an approximation matrix $\mathbf{A}_k$, updated after each iteration k, to take into account of the additional information gain obtained. The cyclic updates make use of the gradient changes, which at each step k provide information about the second derivative of the error function f_k along the optimization search direction. More rigorously, given x_k a partial solution to the optimization problem at the iteration k (we want to converge to the optimal solution x^*), when x_k and x_{k+1} lie near the optimal solution x^*, within which $H(x)$ is positive definite, we can write

$$\mathbf{H}_k\,(x_{k+1} - x_k) = \nabla^2 f_k\,(x_{k+1} - x_k) \sim \nabla f_{k+1} - \nabla f_k.$$

The quasi-Newton method chooses the Hessian approximation $\mathbf{A}_{k+1}$ so that it can well represent the true Hessian. In other words, we require to follow the well-known secant equation condition:

$$\mathbf{A}_{k+1}\,(x_{k+1} - x_k) = \nabla f_{k+1} - \nabla f_k.$$

For completeness, we recall also that the previous equation is intrinsically defined under additional conditions, such as symmetry (typically assumed by the exact Hessian) and the low rank of the difference between successive approximations $\mathbf{A}_k$ and $\mathbf{A}_{k+1}$. In the MLPQNA model, we apply the Hessian approximation known as the BFGS formula, named after its discoverers (Broyden 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970). This is defined by the following equation. Let $\mathrm{sol}_k = x_{(k+1)} - x_k$ and $g_k = \nabla f_{(k+1)} - \nabla f_k$ be the respective matrix terms of equation (2), then we obtain the following rank-2 matrix:

$$\mathbf{A}_{k+1} = \mathbf{A}_k - \frac{\mathbf{A}_k\,\mathrm{sol}_k^T\,\mathbf{A}_k}{\mathrm{sol}_k^T\,\mathbf{A}_k\,\mathrm{sol}_k} + \frac{g_k g_k^T}{g_k^T\,\mathrm{sol}_k}.$$

The BFGS formula generates positive definite approximation matrices under the condition that the initial approximation matrix $\mathbf{A}_0$ is positive definite and the term $g_k^T\,\mathrm{sol}_k > 0$ L-BFGS. From a computational point of view, the BFGS formula is time consuming and requires storing at each step a dense $N \times N$ approximation matrix. Dealing with massive data optimization problems, in order to overcome such requirements, we decided to implement a limited-memory algorithm, known as L-BFGS (Zhu et al. 1997).

The L-BFGS stores at each step only few vectors of length n that represent the approximations implicitly. Despite this improvement in the storage requirements, it yields an acceptable (almost linear) rate of convergence. The main idea of this method is to use error function curvature information from only the most recent iterations to construct the Hessian approximation. Of course, the final result

will not be the Hessian itself, but just an approximation. Surprisingly enough, while the convergence slows down, performances are not affected much and may even improve since it depends on the number of processor's time units spent to calculate the result.

APPENDIX B: SET-UP OF THE EXPERIMENTS

In the following sections, the features are referred to the cardinal number (feature 1: MAG_ISO, etc). For each model, we choose the configuration parameters in order to perform the best results.

B1 Multilayer perceptron trained by back propagation (MLP-BP)

(i) Input nodes (equivalent to the number of features considered in the data set patterns). Max number: 11 (complete patterns); min number: 4 (pruning on optical features); nominal number: 7 (complete optical data set).

(ii) Hidden nodes (depending on the number of features considered in the data set patterns). Max number: 23 (with input nodes in [8, 11]); min number: 15 (with input nodes in [4, 7]).

(iii) Output nodes (based on crispy classification): 2 (1 0 GC, 0 1 not GC).

(iv) Activation functions (neuron function type, used to provide its output, by processing inputs). Input layer: (no input processing, just propagate it); hidden layer: non-linear hyperbolic tangent of input; output layer: linear with softmax normalization (outputs sums up to 1.0 and converge to posterior probabilities).

(v) Learning rule parameters. Output error type: cross-entropy; training mode: batch (weights update after each whole data set patterns calculation); training rule: back propagation with conjugated descent gradient; error loop threshold: 0.001 (one of the stopping criteria); number of iterations: 10 000 (one of the stopping criteria).

B2 Support vector machines (SVM)

(i) Model: C-support vector classification (C-SVC); Kernel: radial basis function.

(ii) Gamma (for each experiment we have a multiplicative step). Min number: 2^{-15}; max number: 2^{23}; step: 4 (multiplicative). C (for each experiment we have a multiplicative step). Min number: 2^{-5}; max number: 2^{15}; step: 4 (multiplicative).

(iii) Error tolerance: 0.001.

(iv) Cache: 100 MB.

(v) Shrinking: on.

(vi) Probability estimates: off.

(vii) Cross validation: k-fold ($k = 5$).

(viii) Weights: 1.

B3 Genetic algorithm model experiment (GAME)

(i) Model: GA with fitness based on trigonometric polynomial expansion.

(ii) Topology: population of chromosomes, each of them composed by genes.

(iii) Input features (depending on the number of features considered in the data set patterns). Max number: 11 (complete data set); min number: 4 (pruning on optical features); nominal number: 7 (complete optical data set).

(iv) Genetic population size (depending on the number of features and polynomial order). Max number: 67 (with 11 features); min number: 25 (with four features).

(v) Population size: $(\text{polynomial}_{\text{order}} * \text{num}_{\text{features}}) + 1$.

(vi) Genetic chromosome size (depending on the polynomial order). Number: 13 (with polynomial order = 6); chromosome size: $(2 \times \text{polynomial}_{\text{order}}) + 1$.

(vii) Output (based on crispy classification). Number in BoK: 1 (0 if no GC; 1 else).

(viii) Output error type: Thresholded MSE (TMSE) with threshold 0.4.

(ix) Error loop threshold: 0.001 (one of the stopping criteria).

(x) Polynomial order: 6.

(xi) Tournament selection (based on the wheel roulette, max probability on the entire population fitness). Number of tournament chromosomes: 2.

(xii) Genetic operators. Crossover probability: 0.9; mutation probability: 0.2; elitism factor: 2.

(xiii) Initial population distribution: Gaussian standard, with all values generated into range $[-1, +1]$.

(xiv) Number of iterations: 10 000 (one of the stopping criteria).

B4 Multilayer perceptron trained by quasi-Newton (MLPQNA)

(i) Input nodes (depending on the number of features considered in the data set patterns). Max number: 11 (complete data set); min number: 4 (pruning on optical features); nominal number: 7 (complete optical data set).

(ii) Hidden nodes (depending on the number of features considered in the data set patterns). Max number: 23 (with input nodes in [8, 11]); min number: 15 (with input nodes in [4, 7]).

(iii) Output nodes (based on crispy classification): number in BoK: 1 (0 if no GC; 1 else).

(iv) Activation functions (neuron function type used to provide its output, by processing inputs). Input layer: no input processing, just propagate it; hidden layer: not linear hyperbolic tangent of input; output layer: linear with softmax normalization (outputs sum up to 1.0 and converge to posterior probabilities).

Learning rule parameters:

(i) output error type: cross-entropy;

(ii) training mode: batch (weights update after each whole data set patterns calculation);

(iii) training rule: quasi-Newton (inverse Hessian approximation by error function gradients);

(iv) QNA implementation rule: based on L-BCFG method (L is for limited memory);

(v) QNA parameters. Decay: 0.001 (weight decay during gradient approximation); restarts: 20 (random restarts for each approximation step); Wstep: 0.01 (stopping threshold, min error for each step); MaxIts: 1500 (max number of Iterations for each approximate step).

B5 Multilayer perceptron trained by genetic algorithms (MLPGA)

(i) Input nodes (depending on the number of features considered in the data set patterns). Max number: 11 (complete data set); min number: 4 (pruning on optical features); nominal number: 7 (complete optical data set).

(ii) Hidden nodes (depending on the number of features considered in the data set patterns). Max number: 23 (with input nodes in [8, 11]); min number: 15 (with input nodes in [4, 7]).

(iii) Output nodes (based on crispy classification). Number in BoK: 1 (0 if no GC; 1 else).

(iv) Activation functions (neuron function type used to provide its output, by processing inputs). Input layer: no input processing, just propagate it; hidden layer: non-linear hyperbolic tangent of input; output layer: non-linear hyperbolic tangent of input.

(v) Learning rule parameters. Output error type: MSE; training mode: batch (weights update after each whole data set patterns calculation); training rule: GA with roulette wheel selection function and fitness based on the MSE between target and output of data set patterns.

(vi) MLPGA parameters. Genetic population size: 25; genetic chromosome size: 13; error loop threshold: 0.001; tournament selection: based on the wheel roulette method (max probability on the entire population fitness); number of tournament chromosomes: 2; crossover probability: 0.9; mutation probability: 0.2; elitism factor: 2; initial population distribution: Gaussian standard, with all values generated into range $[-1, +1]$; number of iterations: 10 000 (one of the stopping criteria).

This paper has been typeset from a TEX/LATEX file prepared by the author.